Single-site monthly streamflow simulation using entropy theory

Z. Hao[1] and V. P. Singh[1,2]

Received 4 November 2010; revised 28 July 2011; accepted 1 August 2011; published 27 September 2011.

[1] Using entropy theory, a new method for single-site monthly streamflow simulation is developed, which is capable of preserving mean, standard deviation, skewness, and lag-one correlation. The method entails deriving joint and conditional probability density functions using the entropy theory, determining the Lagrange multipliers using the information obtained from the historical record, and then simulating streamflow by sequential sampling from the conditional distribution. The advantage of the entropy-based method is that it does not make any assumptions about the probability distribution of the streamflow data. It can also preserve the cross-correlation between streamflow of the last month of the previous year and the first month of the current year and avoid the generation of negative values. The method can also be extended to incorporate higher-order moments and more lag correlations if needed. The disadvantage is that the method will be cumbersome when more statistics need to be preserved and the bimodality that may exist in the empirical probability density function cannot yet be resolved. Application to the Colorado River basin shows that the entropy-based method satisfactorily simulates monthly streamflow.

Citation: Hao, Z., and V. P. Singh (2011), Single-site monthly streamflow simulation using entropy theory, *Water Resour. Res., 47*, W09528, doi:10.1029/2010WR010208.

1. Introduction

[2] Streamflow simulation plays an important role in water resources planning and management. The key requirement for streamflow simulation is that synthetic streamflow sequences preserve key statistical properties of the historical record, such as mean, standard deviation, skewness, and lag correlations. A number of models for streamflow simulation have been proposed, and these models can be classified into two groups: parametric and nonparametric.

[3] A commonly used parametric model for synthetic streamflow generation is the autoregressive moving average (ARMA) model [*Lettenmaier and Burges*, 1977; *Hipel and McLeod*, 1978; *Hipel et al.*, 1979; *Salas and Delleur*, 1980; *Loucks et al.*, 1981; *Vogel and Stedinger*, 1988; *Savic et al.*, 1989], which is quite flexible and can be used for annual as well as seasonal streamflow simulation. The ARMA model is based on the Gaussian assumption which is not usually satisfied by streamflow data. An alternative to the ARMA model for simulating seasonal streamflow is the disaggregation model which has been widely applied [*Valencia and Schaake*, 1973; *Mejia and Rousselle*, 1976]. For the disaggregation model, annual or aggregated streamflow is generated with an appropriate model, and then the generated streamflow is disaggregated to obtain monthly or seasonal streamflow. The disaggregation model ensures the sum of low time scale streamflow values (e.g., monthly) adds up to high time scale streamflow values (e.g., yearly) but has many parameters that need to be estimated. To reduce the number of parameters, several parsimonious models have been proposed, such as a condensed disaggregation model [*Stedinger et al.*, 1985; *Grygier and Stedinger*, 1988] and a stepwise disaggregation model [*Santos and Salas*, 1992]. *Koutsoyiannis and Manetas* [1996] proposed a simple disaggregation model that combines models of lower scale (e.g., monthly) and higher scale (e.g., yearly) with the accurate adjusting procedure.

[4] Parametric models generally require the assumption regarding the marginal distribution of underlying streamflow data. However, the Gaussian assumption usually made may not hold in reality. Therefore, transformation techniques to render the data to be normal are often applied, which in turn give rise to several potential drawbacks. First, some bias of the statistical properties in the original domain may be caused when data is simulated in the transformed domain. Second, negative values may be generated. Third, non-Gaussian features, such as skewness and bimodal, cannot be captured and reproduced efficiently [*Prairie et al.*, 2006]. The autoregressive model with gamma distribution has been proposed to avoid the data transformation [*Fernandez and Salas*, 1990], though the bimodal property cannot be reproduced. Furthermore, it is hard for a usual parametric model to capture the nonlinear relationships that may be observed in the historical record [*Salas and Lee*, 2010].

[5] An attractive alternative is nonparametric models, and *Lall* [1995] provided a review of the application of nonparametric models in hydrology. Nonparametric models are often based on bootstrap techniques or kernel density estimation, and they avoid model selection, minimize (or avoid) parameter estimation, and do not make any assumption about the probability distribution. *Lall and Sharma* [1996] proposed a nearest neighbor bootstrap method for

[1]Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA.
[2]Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas, USA.

0043-1397/11/2010WR010208

resampling monthly streamflow, while probabilistically preserving the dependence structure. To reproduce the serial correlation of historical data, *Vogel and Shallcross* [1996] suggested the moving block bootstrap (MBB) by resampling the observed time series in approximately independent blocks, and compared the method with parametric methods for generating annual streamflow series. *Sharma et al.* [1997] proposed a nonparametric method for monthly streamflow simulation applying the conditional density function with Gaussian kernel, and *Sharma and O'Neill* [2002] extended that method to impose a long-term dependence in the simulated streamflow by incorporating an aggregated variable (denoted as NPL model). *Salas and Lee* [2010] developed a nonparametric method using the K-Nearest Neighbor (KNN) resampling technique with gamma kernel perturbation that can generate data different from the historical record for single site seasonal streamflow simulation. For this method, two approaches, one with the aggregate variable (denoted as KGKA model) and another with the pilot variable (denoted as the KGKP model), were developed to preserve the annual variability.

[6] Nonparametric methods have also been applied for seasonal streamflow simulation with a disaggregation approach. *Tarboton et al.* [1998] developed a nonparametric disaggregation model for simulation based on the conditional distribution obtained by a kernel density estimation method. To address the issue of inefficiency of the kernel density estimation method in higher dimensions, *Prairie et al.* [2007] applied a fast KNN-based bootstrap approach to construct and simulate from the conditional distribution. *Lee et al.* [2010] proposed a space-time disaggregation model based on KNN coupled with a genetic algorithm that can overcome the shortcomings of the models proposed by *Prairie et al.* [2007] and *Koutsoyiannis and Manetas*, [1996]. Based on KNN resampling, *Nowak* [2010] proposed a space-time disaggregation algorithm for disaggregating annual flow to daily flows at different sites.

[7] To simulate streamflow, an assumption about the marginal distribution is often made, especially for parametric models. However, many streamflow records cannot be characterized by commonly assumed probability distributions [*Sharma and O'Neill*, 2002]. The ability to preserve the cross boundary relation (e.g., the correlation between the last season of the previous year and the first season of the current year) and the generation of negative values are two issues that emerge for both parametric and nonparametric models [*Lee et al.*, 2010]. To address the first issue, *Mejia and Rousselle* [1976] made a modification to link past and present values being disaggregated. A practical way to address this problem is to start the generation from a season where the correlation is small. However, this does not work when all correlations between seasons are high. The issue of negative values arises due to the use of normal transformation in parametric models and the application of the Gaussian kernel in nonparametric models. Generally, negative values generated during simulation can be disregarded. However, this solution may not be appropriate when too many negative values are generated.

[8] This study proposes a new model for simulating monthly streamflow at a single site which is capable of preserving key statistics, such as mean, standard deviation, skewness and lag-one correlation. The model is based on entropy theory, wherein a probability distribution function (PDF) is derived without the assumption of normality or the use of a normal transformation. Moreover, the model can preserve the cross-correlation and avoids generation of negative values. It can also be extended to incorporate higher-order moments and more lag correlations if needed. With the specified statistical properties, such as mean, standard deviation, skewness, and lag-one correlation as constraints, the joint probability density function of streamflow of two adjacent months is constructed by maximizing entropy, and the conditional density function is derived from the joint PDF. Then, streamflow is generated using the conditional PDF.

[9] The paper is organized as follows. Describing the framework of the method in section 2, the proposed method is tested using a synthetic example with known underlying model in section 3, followed by an application to the Colorado River basin for streamflow simulation in section 4. Conclusions along with a summary of the main features of the proposed method are given in section 5.

2. Entropy Theory-Based Streamflow Simulation

[10] The first step in the streamflow simulation is the derivation of joint and conditional probability density functions of streamflow. The derivation involves the expression of the joint Shannon entropy, specification of constraints based on the statistics to be preserved, maximization of the entropy subject to the specified constraints, and determination of the Lagrange multipliers. Then the monthly streamflow is simulated from the conditional distribution sequentially.

2.1. Shannon Entropy

[11] For a continuous random variable $X \varepsilon\ [a, b]$ with a probability density function (PDF) $f(x)$, the Shannon entropy E_1 is defined as [*Shannon*, 1948; *Shannon and Weaver*, 1949]:

$$E_1 = -\int_a^b f(x)\,\ln f(x)\,dx, \tag{1}$$

where x is a value of random variable X with lower limit a and upper limit b. For a bivariate case involving two continuous random variables X and Y or random vector (X, Y) with joint probability density function $f(x, y)$ defined over the space $[a, b] \times [c, d]$, the Shannon entropy can be defined as

$$E_2 = -\int_c^d \int_a^b f(x,y)\,\ln f(x,y)\,dxdy. \tag{2}$$

2.2. Specification of Constraints

[12] For streamflow simulation, it is desired to preserve such statistics as mean, standard deviation, skewness, and lag-one correlation. These statistics can be regarded as constraints for deriving the distribution of streamflow. Then, sampling from the distribution can be expected to preserve these required statistics. The mean, standard deviation, and skewness of streamflows can be determined through the first three moments. In order to preserve the correlation between streamflows of two adjacent months (say, January and February), the joint PDF of the continuous random

vector (X, Y) is needed for which constraints in general form can be stated as

$$\int_c^d \int_a^b g_i(x,y) f(x,y) dxdy = \overline{g_i} \quad i = 0, 1, 2, \ldots, m. \tag{3}$$

[13] Specifically,

$$\int_c^d \int_a^b f(x,y) dxdy = 1 \quad i = 0, \tag{4a}$$

$$\int_c^d \int_a^b x^i f(x,y) dxdy = \overline{x^i} \quad i = 1, 2, 3, \tag{4b}$$

$$\int_c^d \int_a^b y^{i-3} f(x,y) dxdy = \overline{y^{i-3}} \quad i = 4, 5, 6, \tag{4c}$$

$$\int_c^d \int_a^b xy f(x,y) dxdy = \overline{xy} \quad i = 7, \tag{4d}$$

where x and y are streamflow values of adjacent months; $g_i(x, y)$ (or g_i) is a known function of random vector (X, Y), which can be specified as $g_0 = 1$, $g_1 = x$, $g_2 = x^2$, $g_3 = x^3$, $g_4 = y$, $g_5 = y^2$, $g_6 = y^3$ and $g_7 = xy$ for the proposed constraints; $\overline{g_i}$ is the expected value of the function $g_i(x, y)$ (e.g., if $g_1(x, y) = x$, then $\overline{x}$ is the mean of X) estimated from the historical record; $\overline{x^i}$ and $\overline{y^{i-3}}$ (i = 1, 2, ..., 7) are the first to third noncentral moments of random variables X and Y, respectively; $\overline{xy}$ is the expectation of XY and m is the number of constraints ($m = 7$ in this case). The constraint in equation (4a) assures that the integration of the probability density function over the whole interval should be unity, which is often termed as the "normalization condition" or the "total probability theorem."

2.3. Maximization of Entropy and Derivation of Probability Distributions

[14] According to the principle of maximum entropy, formulated by *Jaynes* [1957], the least biased probability distribution will be the one that maximizes the Shannon entropy subject to the given constraints. To derive the joint PDF of streamflows of two adjacent months (say, January and February), the entropy given by equation (2) is maximized, subject to the constraints given by equation (3). The maximization can be performed using the method of Lagrange multipliers.

[15] Denoting the Lagrange multipliers for the joint PDF of January and February streamflows as $\mathbf{\Phi}_{1,2} = [\lambda_0, \lambda_1, \ldots, \lambda_7]$, where $\lambda_0, \lambda_1, \ldots, \lambda_7$ are the Lagrange multipliers, the Lagrangian function L, using equations (2) and (3) can be expressed as [*Kapur*, 1989]

$$\begin{aligned} L = &-\int_c^d \int_a^b f(x,y) \ln f(x,y) dxdy - (\lambda_0 - 1) \\ &\cdot \left[\int_c^d \int_a^b f(x,y) dxdy - 1\right] \\ &- \sum_{i=1}^m \lambda_i \left[\int_c^d \int_a^b g_i(x,y) f(x,y) dxdy - \overline{g_i}\right]. \end{aligned} \tag{5}$$

[16] Differentiating L with respect to f and setting the derivative to 0, the maximum entropy-based joint probability density function is obtained with representation of $g_i(x, y)$ by their specific values as [*Kesavan and Kapur*, 1992]

$$\begin{aligned} f(x,y) &= \exp\left[-\sum_{i=0}^m \lambda_i g_i(x,y)\right] \\ &= \exp\left(-\lambda_0 - \sum_{i=1}^3 \lambda_i x^i - \sum_{i=4}^6 \lambda_i y^{i-3} - \lambda_7 xy\right). \end{aligned} \tag{6}$$

[17] Substituting equation (6) in the "normalization condition" in equation (4a), one can obtain the zeroth Lagrangre multiplier λ_0 as a function of other Lagrange multipliers as

$$\exp(\lambda_0) = \int_c^d \int_a^b \exp\left[-\sum_{i=1}^m \lambda_i g_i(x,y)\right] dxdy. \tag{7}$$

[18] The joint PDF given by equation (6) has unknown Lagrange multipliers, λ_i (i = 1, ..., 7), that need to be determined.

[19] For monthly streamflow simulation, 12 joint density functions with random vector ($X_{t,n}$, $Y_{t,n}$) have to be estimated from the historical data, where t is the year and n (n = 1, 2, ..., 12) is the month. For the joint distribution of December and January streamflows, the random vector has to be replaced by ($X_{t-1,12}$, $Y_{t,1}$) to preserve the cross-correlation between streamflow in December of the previous year ($X_{t-1,12}$) and that in January of the current year ($Y_{t,1}$).

[20] The marginal density function for X can be obtained by integrating the joint PDF $f(x, y)$ given by equation (6) over Y as

$$\begin{aligned} f(x) &= \int_c^d f(x,y) dy \\ &= \exp(-\lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3 - \lambda_0) \\ &\quad \cdot \int_c^d \exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy) dy. \end{aligned} \tag{8}$$

[21] The conditional density function of Y given $X = x$ can now be obtained as

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy)}{\int_c^d \exp(-\lambda_4 y - \lambda_5 y^2 - \lambda_6 y^3 - \lambda_7 xy) dy}. \tag{9}$$

[22] The conditional cumulative distribution function $F_{Y|X}(y|x)$ of Y given $X = x$ can be written as

$$F_{Y|X}(y|x) = \int_c^y f(z|x) dz. \tag{10}$$

2.4. Parameter Estimation

[23] The Lagrange multipliers contained in equation (6) are now determined. Substitution of equation (6) in equation (3) results in a set of nonlinear equations whose solution results in the Lagrange multipliers:

$$\begin{aligned} &\int_c^d \int_a^b g_i(x,y) \exp\left[-\sum_{k=0}^m \lambda_k g_k(x,y)\right] dxdy = \overline{g_i} \\ &i = 0, 1, 2, \ldots, m. \end{aligned} \tag{11}$$

[24] In general, an analytical solution for obtaining the Lagrange multipliers (for $m > 2$) does not exist and numerical solution is the only resort. It has been shown that the problem of solving the set of nonlinear equations is equivalent to finding the minimum of a convex function Γ expressed as [*Mead and Papanicolaou*, 1984; *Kapur*, 1989]

$$\begin{aligned}\Gamma &= \lambda_0 + \sum_{i=1}^{m} \lambda_i \overline{g_i} \\ &= \ln \int_c^d \int_a^b \exp\left[-\sum_{i=1}^{m} \lambda_i g_i(x,y)\right] dxdy + \sum_{i=1}^{m} \lambda_i \overline{g_i}.\end{aligned} \tag{12}$$

[25] The Newton-Raphson method can be applied to achieve the minimization of the convex function yielding the Lagrange multipliers $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_7]'$ as follows. Starting from some initial value $\boldsymbol{\lambda}_{(0)}$, one updates $\boldsymbol{\lambda}_{(1)}$ using the equation

$$\boldsymbol{\lambda}_{(1)} = \boldsymbol{\lambda}_{(0)} - H^{-1} \frac{\partial \Gamma}{\partial \lambda_i} \quad i = 1, 2, \ldots, 7, \tag{13}$$

where the gradient of the convex function is expressed as

$$\frac{\partial \Gamma}{\partial \lambda_i} = \overline{g_i} - \int_c^d \int_a^b \exp\left[-\sum_{k=0}^{m} \lambda_k g_k(x,y)\right] g_i(x,y) dxdy \qquad i = 1, 2, \ldots, 7,$$

and the Hessian matrix $\mathbf{H}$ is expressed as

$$H = \begin{bmatrix} \mathrm{var}(x) & \mathrm{cov}(x, x^2) & \ldots & \mathrm{cov}(x, xy) \\ \mathrm{cov}(x^2, x) & \mathrm{var}(x^2) & \ldots & \mathrm{cov}(x^2, xy) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(xy, x) & \mathrm{cov}(xy, x^2) & \ldots & \mathrm{var}(xy) \end{bmatrix}$$

where elements of the Hessian matrix are expressed as

$$\begin{aligned}H_{i,j} &= \mathrm{cov}(g_i, g_j) \\ &= \int_c^d \int_a^b \exp\left[-\sum_{k=0}^{m} \lambda_k g_k(x,y)\right] g_i(x,y) g_j(x,y) dxdy, \\ &\quad - \int_c^d \int_a^b \exp\left[-\sum_{k=0}^{m} \lambda_k g_k(x,y)\right] g_i(x,y) dxdy \\ &\quad \cdot \int_c^d \int_a^b \exp\left[-\sum_{k=0}^{m} \lambda_k g_k(x,y)\right] g_j(x,y) dxdy \\ &\quad i,j = 1, 2, \ldots, m,\end{aligned}$$

where $\mathrm{cov}(g_i,\ g_i) = \mathrm{var}(g_i)$ and $\mathbf{H}^{-1}$ is the inverse of Hessian matrix $\mathbf{H}$. In this study, the MATLAB function *fminsearch* was used to obtain the minimum of equation (12) and hence the Lagrange multipliers.

[26] For the generation of monthly streamflow, 12 joint PDFs of streamflows of two adjacent months are needed and the corresponding Lagrange multiplier sets ($\boldsymbol{\Phi}_{1,2}$, $\boldsymbol{\Phi}_{2,3}, \ldots, \boldsymbol{\Phi}_{12,1}$) of each joint PDF have to be estimated. Each Lagrange multiplier in the joint PDF in equation (6) is related to one statistic that is to be preserved. For instance, in parameters $\boldsymbol{\Phi}_{1,2} = [\lambda_0, \lambda_1, \ldots, \lambda_7]$ of the joint PDF of streamflow of January and February, λ_1, λ_2, and λ_3 relate to the mean, standard deviation and skewness of the January streamflow, λ_4, λ_5, and λ_6 relate to the mean, standard deviation and skewness of the February streamflow, and λ_7 is the parameter relating to the lag-one correlation of streamflows of two adjacent months. Likewise, parameters $\boldsymbol{\Phi}_{2,3}$ of the joint PDF of streamflow of February and March relate to the required statistics for the February and March streamflows and so on. If more statistics (e.g., kurtosis) need to be preserved, one can incorporate the corresponding Lagrange multipliers in the joint PDF. Thus, the entropy-based formulation is quite flexible and can be extended to incorporate more statistical properties, if needed.

2.5. Generation

[27] There are several techniques that can be employed for the generation of random values from the bivariate distribution, such as the conditional distribution method, the transformation method, the acceptance/rejection method, and the composition method [*Johnson*, 1987; *Balakrishnan and Lai*, 2009]. In order to sample from the continuous joint PDF $f(x, y)$ to obtain the random values of (X, Y), the conditional distribution method was employed in this study. For the generation of streamflow while preserving the correlation between adjacent months, streamflow values of one month can be generated from the conditional distribution given the streamflow value of its previous month. When the method is applied to the generation of monthly data of each year, 12 conditional distributions have to be used sequentially. To illustrate this method, it is assumed that the simulation starts from January and n year data is to be generated. Let $x_{t,s}$ denote streamflow for month s of year t ($s = 1, 2, \ldots, 12$, $t = 1, 2, \ldots, n$). The step by step simulation procedure for generating random values of each month can now be summarized as follows.

[28] (1) Generate a random vector ($x_{1,1}$, $x_{1,2}$) from the joint density function given by equation (6) with parameters $\boldsymbol{\Phi}_{1,2}$ estimated in section 2.4. The random values $x_{1,1}$ and $x_{1,2}$ are the January and February streamflows of the first year.

[29] (2) With the initial value $x_{1,2}$ generated in step (1), one can generate the March streamflow of the first year $x_{1,3}$ from the conditional cumulative distribution function in equation (10) with parameters $\boldsymbol{\Phi}_{2,3}$. To that end, generate a uniform distributed random value w_1 between [0, 1] which can be done with the use of random number generator function *rand* in MATLAB. This w_1 value can be considered to be the conditional cumulative probability corresponding to a specific value $x_{1,3}$, given the initial value $x_{1,2}$. This can be expressed with equation (10) as

$$F_{Y|X}(x_{1,3}|x_{1,2}) = w_1.$$

[30] Then, $x_{1,3}$ can be generated by solving the above equation. Similarly, monthly streamflows $x_{1,4}, \ldots, x_{1,12}$ can be generated while parameters $\boldsymbol{\Phi}_{3,4}, \ldots, \boldsymbol{\Phi}_{11,12}$ are used sequentially. Then monthly streamflow of the first year can be generated.

[31] (3) With $x_{1,12}$ (December streamflow of the first year) generated in step (2), $x_{2,1}$ (January streamflow of the second year) can be generated with the parameter $\mathbf{\Phi}_{12,1}$ similarly. In this manner, monthly streamflow of the second year, $x_{2,1}, \ldots, x_{2,12}$, can be generated.

[32] (4) Repeat step (3) until monthly streamflow of the nth year is generated.

[33] In the above steps, numerical integration is performed to generate random values from the inverse (conditional) cumulative distribution.

3. Testing With Synthetic Data

[34] In order to test the performance of the proposed method to approximate the density function of the known model and reproduce the statistics of samples from it, the bivariate gamma distribution was selected. The gamma marginal distribution of a random variable z with scale parameter β and shape parameter γ is defined as

$$f(z) = \frac{z^{\gamma-1}}{\Gamma(\gamma)\beta^{\gamma}} \exp(-z/\beta). \tag{14}$$

[35] The five-parameter bivariate gamma distribution $f(x, y)$ by *Smith et al.* [1982] can be expressed as

$$f(x,y) = \frac{(\beta_1 x)^{\gamma_1-1}(\beta_2 y)^{\gamma_2-1}\exp[-(\beta_1 x+\beta_2 y)/(1-\eta)]}{(1-\eta)^{\gamma_1}\Gamma(\gamma_1)\Gamma(\gamma_2-\gamma_1)} \cdot \sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\frac{\eta^{j+k}}{(1-\eta)^{2j+k}}\frac{\Gamma(\gamma_2-\gamma_1+k)}{\Gamma(\gamma_2+j+k)}\frac{(\beta_1 x\beta_2 y)^j(\beta_2 y)^k}{j!k!}, \tag{15}$$

where β_1 and β_2 are the scale parameters; γ_1 and γ_2 are the shape parameters; $\eta = \rho(\gamma_2/\gamma_1)^{0.5}$, where ρ is the correlation coefficient between x and y. Parameters of the bivariate gamma distribution were specified as: $\beta_1 = 1$, $\gamma_1 = 6$, $\beta_2 = 5$, $\gamma_1 = 9$, and $\rho = 0.25$.

[36] One sample consisting of 3000 data pairs was drawn from the bivariate gamma distribution, which is regarded as the calibration sample for fitting and evaluating the entropy-based method. To quantify the performance of the proposed method in approximating the marginal and bivariate gamma PDFs, the root mean square error (RMSE) was computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - p_i)^2}, \tag{16}$$

where n is the length of the data, s_i and p_i ($i = 1, \ldots, n$) are the maximum entropy-based probability densities of equations (6) and (8) and densities from the marginal and bivariate gamma distribution of equation (14) and (15) corresponding to the ith value.

[37] The maximum entropy-based marginal PDFs of random variable X and Y in equation (8) estimated from the calibration sample together with the gamma marginal PDFs in equation (14) were plotted, as shown in Figure 1. As can be seen the maximum entropy-based density of X was virtually indistinguishable from that of the gamma density of X. Generally, the maximum entropy-based density approximates the gamma density of Y relatively well, though some discrepancies exist. The RMSE values between the maximum

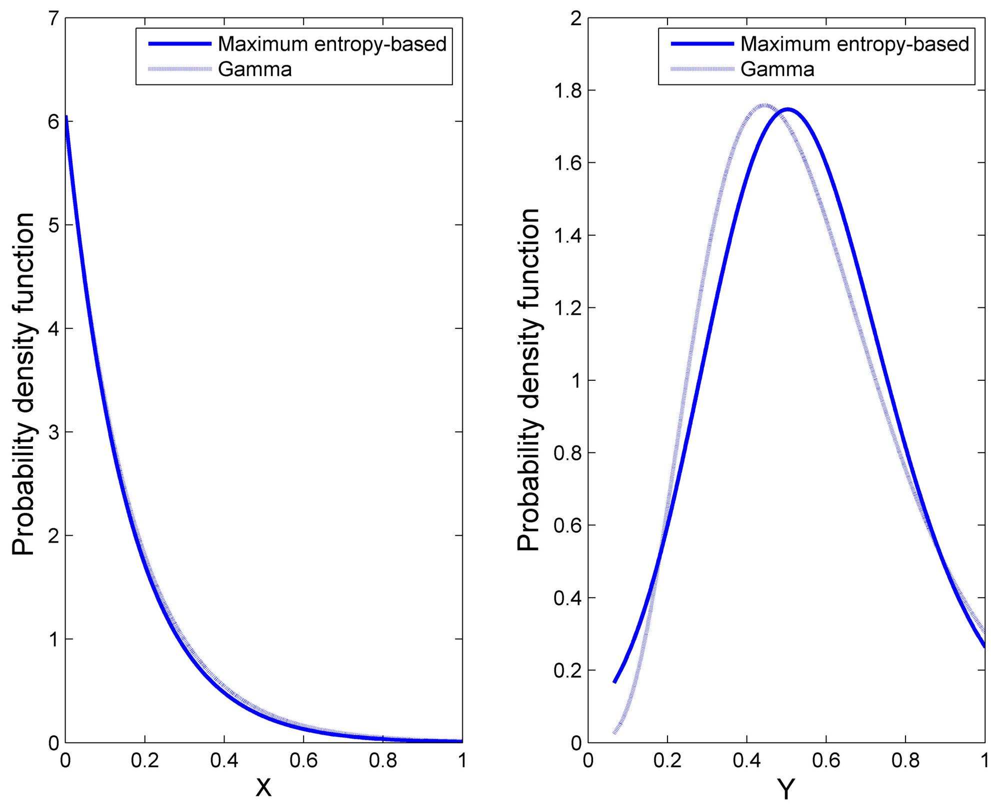

Figure 1. Maximum entropy-based marginal PDFs in equation (8) and gamma marginal PDFs in equation (14) for variables X and Y.

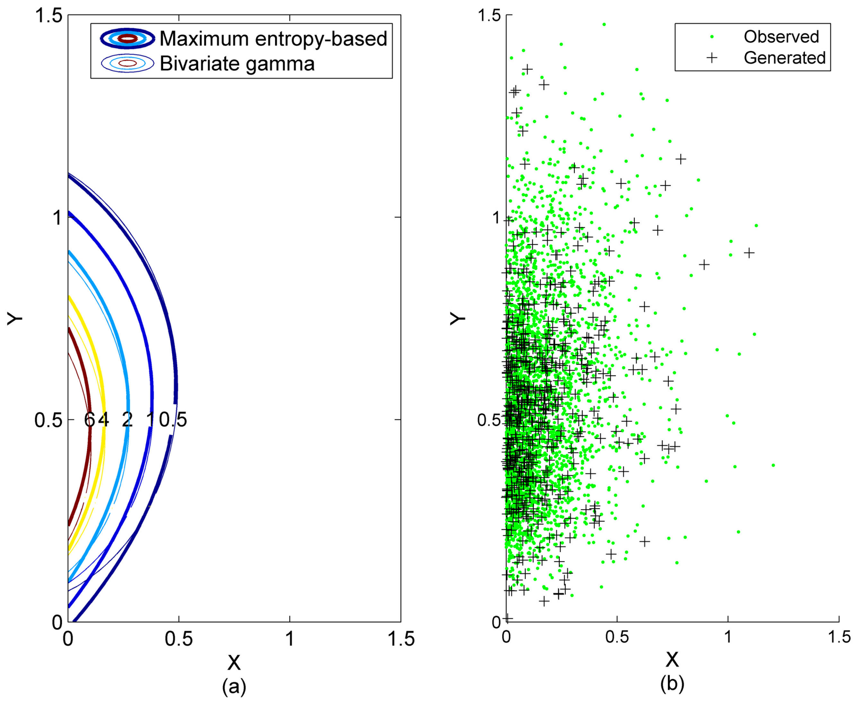

Figure 2. (a) Contours of maximum entropy-based joint PDF in equation (6) and bivariate gamma PDF in equation (15). (b) Comparison of generated data pairs with the calibration sample (or observed data pairs).

entropy-based density and gamma density were 0.062 for variable X and 0.14 for variable Y, respectively. Thus, the maximum entropy-based marginal PDFs estimated from the calibration sample can approximate the gamma marginal PDFs relatively well.

[38] Contours of the maximum entropy-based joint PDF in equation (6) estimated from the calibration sample and the underlying bivariate gamma PDF in equation (15) were plotted, as shown in Figure 2(a). The contour lines of the maximum entropy-based PDF approximate those of the underlying gamma PDF relatively well. The RMSE value between the maximum entropy-based joint density and underlying bivariate gamma density was 0.74. A bivariate sample with 500 data pairs was generated and is shown together with the calibration sample in Figure 2(b). It is seen that generally the spreading pattern of the generated data pairs matches that of the calibration sample well. This shows that the proposed method approximated the underlying bivariate gamma PDF relatively well.

[39] One-hundred bivariate samples each consisting of 3000 data pairs were generated from the maximum entropy-based joint PDF in equation (6) estimated from the calibration sample. Statistics of generated data pairs and the calibration sample were compared using box plots, including the mean, standard deviation, skewness, and lag-one correlation. The central mark of the box is the median and the end lines of the box represent 25th and 75th percentiles. The whiskers are the maximum and minimum values of the simulated statistics. A wide box plot signifies large variability. When a statistic falls in the box plot, the performance is considered to be good [*Prairie et al.*, 2007; *Nowak*, 2010; *Salas and Lee*, 2010]. Statistics of the generated data pairs and the calibration sample were compared with box plots as shown in Figure 3. All statistics fell in the box plots, and this showed that the proposed method can preserve the mean, standard deviation, skewness, and lag-one correlation well.

4. Application to Simulation of Monthly Streamflow

[40] The entropy-based method was applied to monthly streamflow at 10 sites (site 11 to site 20) in the Colorado

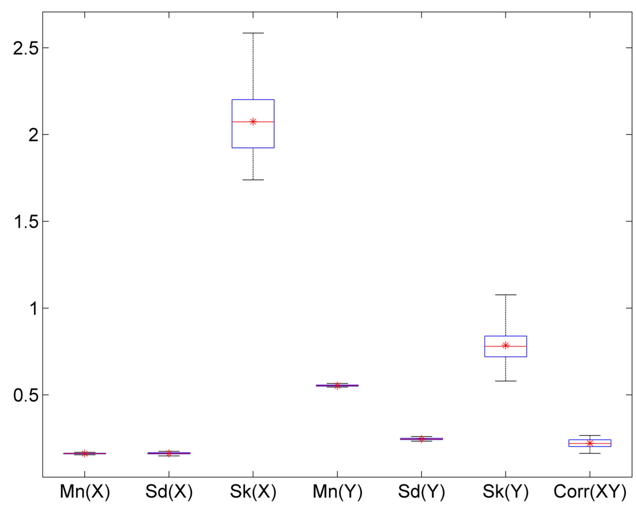

Figure 3. Box plots of statistics of the calibration sample and generated data pairs. Mn, Sd, and Sk represent the mean, standard deviation, and skewness. Corr represents the correlation. Star marks represent statistics of the calibration sample.

River basin from 1906–2003 [*Lee and Salas*, 2006]. (These data can be found at http://www.usbr.gov/lc/region/g4000/NaturalFlow/previous.html). Without loss of generality, the monthly streamflow data was scaled to [0, 1] for computational convenience. For the original data (OD) of each month with maximum value MX and minimum value MN, the scaled data (SD) of each month was expressed as: SD $= [\mathrm{OD} - (1 - d)\mathrm{MN})]/[(1 + d)\mathrm{MX} - (1 - d)\mathrm{MN}]$, where d is a scale parameter, which was selected as 0.05 in this study. With the use of constraints in equations (4a) to (4d), parameters $(\mathbf{\Phi}_{1,2}, \mathbf{\Phi}_{2,3}, \ldots, \mathbf{\Phi}_{12,1})$ of each joint PDF in equation (6) were first estimated. Then, the conditional distribution was derived from the known joint PDF using equations (6) and (8). Thereafter, samples were drawn sequentially using the procedure outlined in section 2.5 and then transformed back to the original domain. From the scaling expression, the values of 0 and 1 in the scaled domain corresponded to the $(1 - d)$MN and $(1 + d)$MX in the original domain, and thus the values outside the observed streamflow range can be generated.

[41] Simulation results were satisfactory for all stations, as shown for the station, Lees Ferry, Arizona, on the Colorado River (U.S. Geological Survey station number 09380000), which has been used in earlier studies [*Prairie et al.*, 2006; *Salas and Lee*, 2010]. One-hundred flow sequences, each of 100 and 400 years long, termed as S_1 and S_2, respectively, were generated to test the proposed method. Statistics of generated and historical data, including the mean, standard deviation, skewness, lag-one correlation, maximum and minimum values, were compared using box plots. Furthermore, other statistics pertaining to low values reflecting drought conditions, such as maximum drought length, maximum drought amount, maximum surplus length, maximum surplus amount, and storage capacity, were also compared for generated and historical data.

[42] The box plots was used to measure the performance of the proposed method and the performance was considered to be good when a statistic fell in the box as described in section 3. To quantify the performance of the entropy-based method, absolute error (AE) and relative error (RE) of the simulated statistics were computed as $\mathrm{AE} = S_m - X_o$ and $\mathrm{RE} = (S_m - X_o)/X_o$, where S_m is the median of simulated statistic for the generated data, and X_o is the statistic for the historical data.

4.1. Validation of Marginal PDF and Joint PDF

[43] Maximum entropy-based marginal PDFs and empirical histograms of scaled streamflows were constructed and compared, as shown for two sample months of May and June in Figure 4. Note that the marginal entropy based PDF in equation (8) of streamflow of a specific month, say May, can either be derived from the joint density function of April and May streamflow with parameter $\mathbf{\Phi}_{4,5}$ (denoted as $f_{5,45}$) or from the joint density function of May and June streamflows with parameter $\mathbf{\Phi}_{5,6}$ (denoted as $f_{5,56}$). Though the PDF of May streamflows can be derived from different joint distributions with different Lagrange multipliers, densities $f_{5,45}$ and $f_{5,56}$ should be close to each other, which is verified in Figure 4. The probability density function of the May streamflow was bimodal, which has been shown by *Prairie et al*. [2006]. The maximum entropy-based PDFs fitted the empirical histograms relatively well, except that the

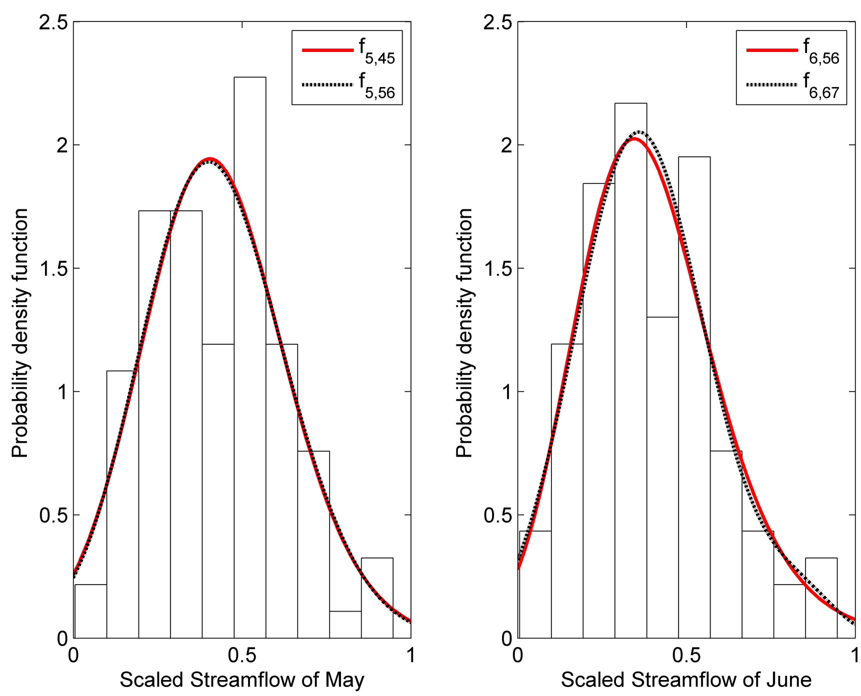

Figure 4. Maximum entropy-based marginal PDFs and empirical histograms for scaled streamflow of May and June of the Colorado River at Lees Ferry ($f_{5,45}$, marginal PDF for May streamflow with parameter $\mathbf{\Phi}_{4,5}$; $f_{5,56}$, marginal PDF for May streamflow with parameter $\mathbf{\Phi}_{5,6}$; $f_{6,56}$, marginal PDF for June streamflow with parameter $\mathbf{\Phi}_{5,6}$; $f_{6,67}$, marginal PDF for June streamflow with parameter $\mathbf{\Phi}_{6,7}$).

bimodality in the density of the scaled May streamflow could not be resolved. Contours of maximum entropy-based and empirical joint densities are shown in Figures 5(a) and 5(b). The historical data spread along the contours as seen in Figure 5(a). The maximum entropy-based joint densities matched the empirical densities well for most parts as shown in Figure 5(b). For example, the maximum entropy-based joint density values near the empirical contour line with a density of 2 were 1.5, 2, and 2.6.

4.2. Monthly Mean, Standard Deviation, Skewness, and Lag-One Correlation

[44] Statistics of generated and historical data for all months of simulation S_1 were computed, as shown in Figure 6, and then the absolute error and relative error for each statistic were calculated, as shown in Figure 7 and Table 1. The median values of simulated mean, standard deviation, skewness, and lag-one correlation were close to those of the historical data. All statistics of mean, standard deviation, skewness, and lag-one correlation fell in the box plots, indicating the goodness of the entropy-based method. Even though the absolute error was relatively large for several months, like that for the mean of May and standard deviation of June, as seen from Figure 7, the result was satisfactory based on the relative error in Table 1. The relative error of mean, standard deviation, and lag-one correlation was lower than 5% and that of skewness was lower than 10% for all months. The relative error of simulated skewness was relatively high and was not preserved as well as other statistics. The lag-one correlation between the December streamflow of the previous year and the January streamflow of the current year was also preserved well, as seen from Table 1 in that the relative error was −0.46%. *Salas and Lee* [2010] showed that nonparametric model with the long-term dependence (NPL) model underestimated the skewness throughout the year and overestimated the standard deviation for wet months, while the KNN resampling technique with gamma kernel perturbation with the aggregate variable (KGKA) and the pilot variable (KGKP) underestimated lag-one correlation. Since all these statistics were preserved well and no underestimation or overestimation existed, the entropy-based method performed better in preserving the four statistics.

[45] Statistics of generated and historical data for all months of simulation S_2 are shown in Figure 8. It is seen that all statistics fell in the box plots, indicating satisfactory model performance for simulation S_2. In addition, the box plots became narrower and thus the variability of simulated statistics was reduced, as shown in Figure 8. Specifically, comparison of statistics of the generated and historical data of selected months, January and May, for simulation S_1 and S_2 with different percentiles is shown in Table 2. For the

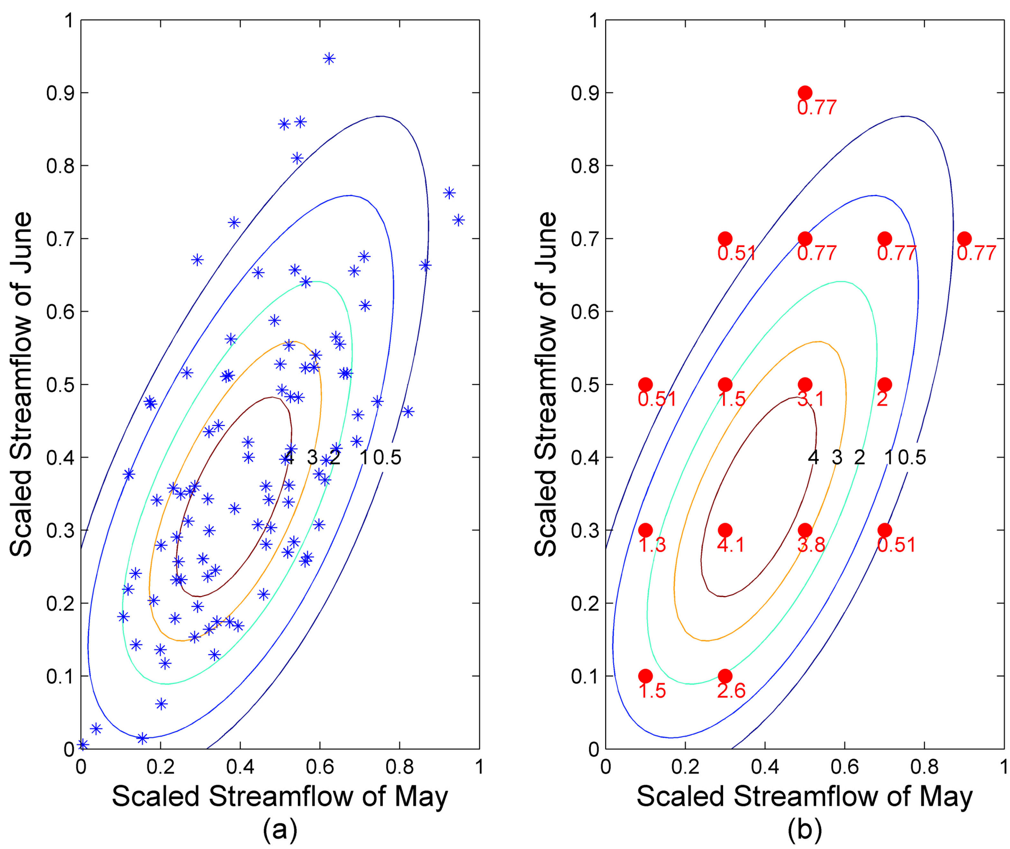

Figure 5. Contours of the maximum entropy-based and empirical joint PDFs of the scaled May and June streamflow with historical data of the Colorado River at Lees Ferry. (a) Contours of maximum entropy-based joint PDF and historical data plotted as stars. (b) Contours of maximum entropy-based joint PDF and empirical joint PDF plotted as points.

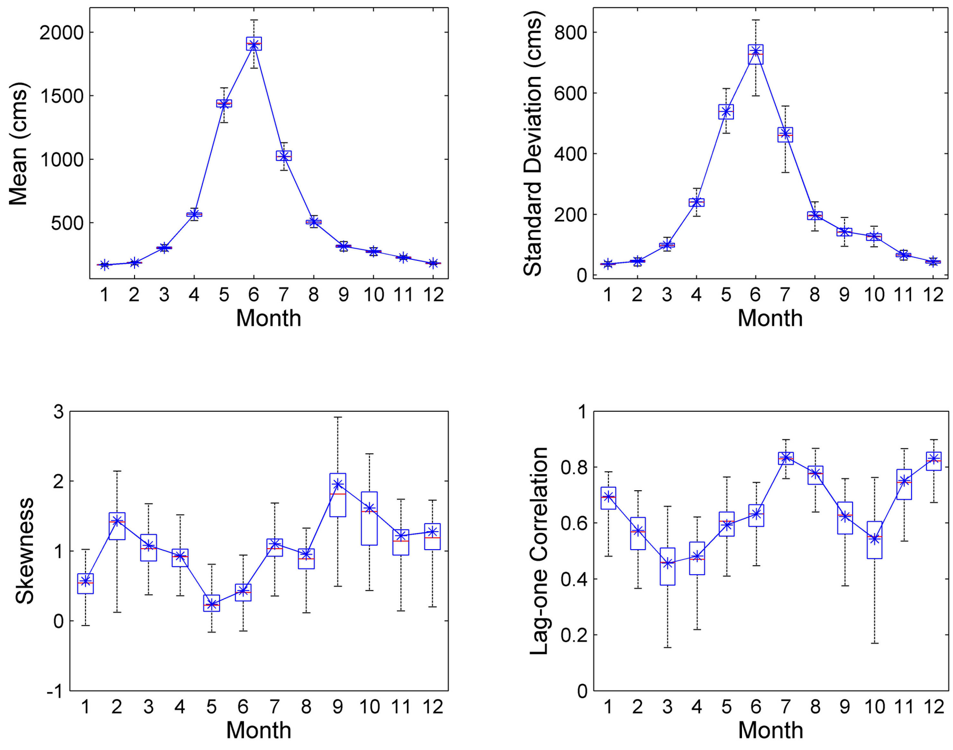

Figure 6. Box plots of mean, standard deviation, skewness, and lag-one correlation of generated and historical data for simulation S_1. Continuous lines with star marks for each month represent statistics of the historical data. The units for mean and standard deviation are in cubic meter per second (cms).

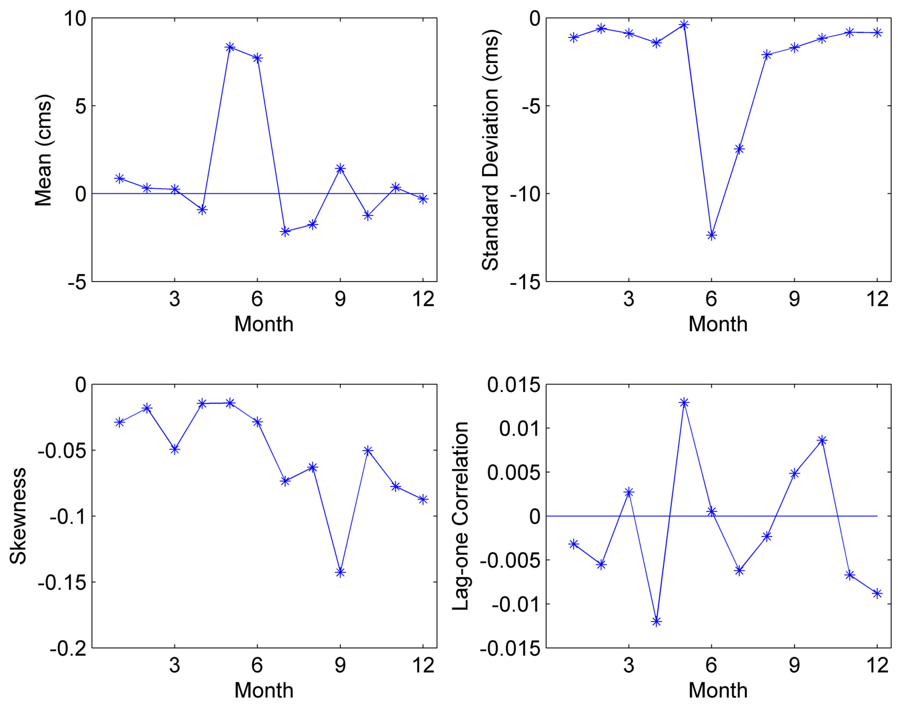

Figure 7. Absolute errors of mean, standard deviation, skewness, and lag-one correlation for simulation S_1. The units for mean and standard deviation are in cubic meter per second (cms).

Table 1. Relative Error (%) in Statistics for Each Month for Simulation S_1[a]

Statistics/Month	1	2	3	4	5	6	7	8	9	10	11	12
Mean	0.52	0.17	0.08	−0.16	0.58	0.41	−0.21	−0.35	0.46	−0.46	0.16	−0.17
SD	−3.04	−1.31	−0.90	−0.59	−0.07	−1.67	−1.59	−1.06	−1.18	−0.91	−1.23	−1.89
Skewness	−5.07	−1.27	−4.57	−1.58	−6.03	−6.60	−6.65	−6.63	−7.30	−3.12	−6.36	−6.86
Lag-one Correlation	−0.46	−0.96	0.60	−2.49	2.18	0.08	−0.74	−0.30	0.78	1.59	−0.89	−1.06

[a]Units for the mean and SD are in cubic meter per second (cms).

simulation of skewness of the May streamflow, 25th, 50th, and 75th percentiles of simulated skewness were 0.15, 0.24, 0.36 in simulation S_1 and 0.19, 0.24, 0.31 in simulation S_2, respectively. The interquartile range (distance between 25th percentile and 75th percentile) is 0.12 in simulation S_2 which was smaller than 0.21 in simulation S_1.

[46] Generally, the median values of statistics of simulated data matched those of historical data well when the length of generated annual streamflow was longer than 100 and simulation of these statistics could be further improved by generating a longer record. Nevertheless, for a streamflow record with a length of annual streamflow around 100, the proposed method satisfactorily preserved the mean, standard deviation, skewness, and lag-one correlation.

4.3. Monthly Maximum and Minimum Values

[47] The maximum and minimum values of generated data and historical data for all months of simulation S_1 and S_2 were obtained, as shown in Figure 9. There was no significant overestimation or underestimation of the maximum and minimum values for most months of simulation S_1. However, for many months of simulation S_2, the maximum values were overestimated and the minimum values were underestimated. The value of scale parameter d affects the generated maximum and minimum values and when d equates zero the generated values are bounded by the observed maximum and minimum values. In both simulation S_1 and S_2, no negative values were generated.

[48] *Salas and Lee* [2010] showed that the nonparametric NPL model preserved the maximum values well, although the minimum values were underestimated, whereas both the KGGA and KGKP models preserved the maximum and minimum values well. Since the overestimation of maximum values and underestimation of minimum values occurred when a relatively long record of annual streamflow were generated, the proposed method did not perform as well.

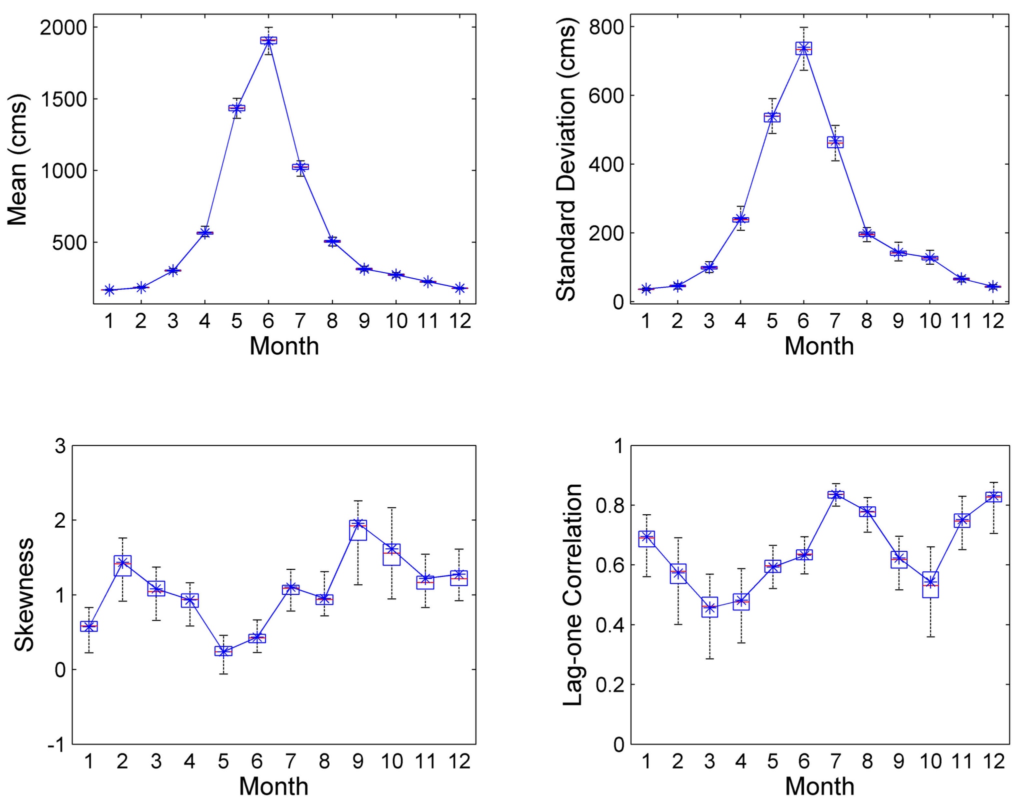

Figure 8. Box plots of mean, standard deviation, skewness, and lag-one correlation of generated and historical data for simulation S_2. Continuous lines with star marks for each month represent statistics of the historical data. The units for mean and standard deviation are in cubic meter per second (cms).

Table 2. Comparison of Statistics of Generated and Observed Streamflows of Jan and May for Simulation S_1 and S_2[a]

Statistics	Simulation S_1	Simulation S_2	Percentile	Observation
		January		
Mean	165	166	25th	167
Mean	167	168	50th	167
Mean	170	169	75th	167
SD	33	35	25th	37
SD	36	36	50th	37
SD	38	37	75th	37
Skewness	0.39	0.51	25th	0.57
Skewness	0.58	0.58	50th	0.57
Skewness	0.69	0.64	75th	0.57
Lag-one Correlation	0.63	0.66	25th	0.70
Lag-one Correlation	0.68	0.69	50th	0.70
Lag-one Correlation	0.72	0.71	75th	0.70
		May		
Mean	1408	1416	25th	1433
Mean	1437	1436	50th	1433
Mean	1472	1454	75th	1433
SD	512	523	25th	539
SD	538	540	50th	539
SD	560	548	75th	539
Skewness	0.15	0.19	25th	0.24
Skewness	0.24	0.24	50th	0.24
Skewness	0.36	0.31	75th	0.24
Lag-one Correlation	0.55	0.58	25th	0.59
Lag-one Correlation	0.59	0.60	50th	0.59
Lag-one Correlation	0.64	0.62	75th	0.59

[a]Units for the mean and SD are in cubic meter per second (cms).

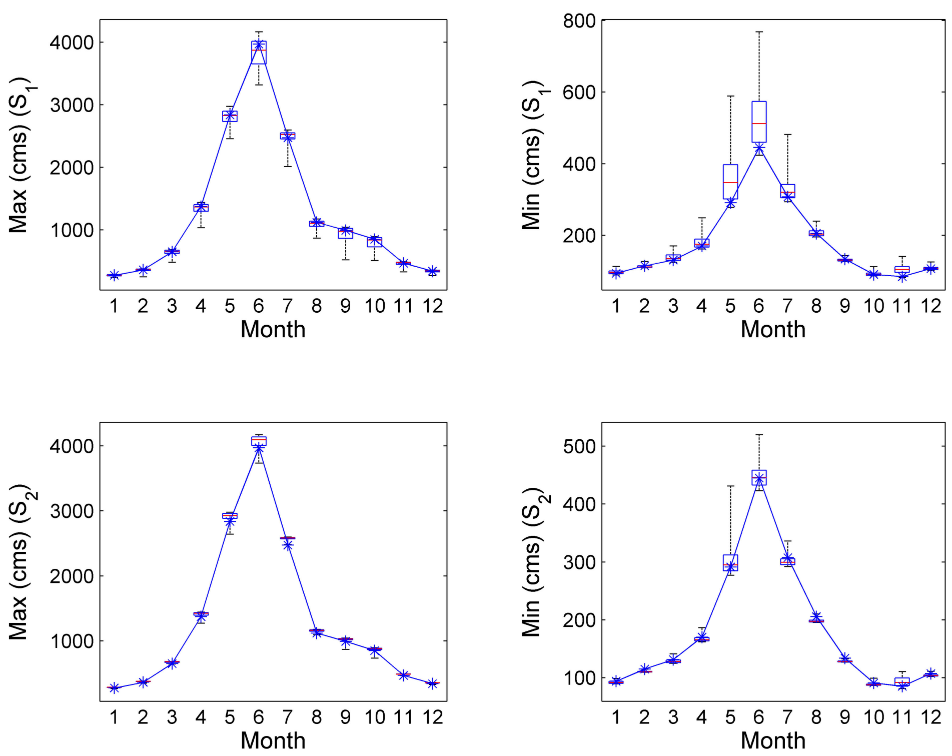

Figure 9. Box plots of maximum and minimum values of generated and historical data for simulation S_1 and S_2. Continuous lines with star marks for each month represent historical data. The units for maximum and minimum values are in cubic meter per second (cms).

4.4. Extension to Higher-Order Moments

[49] The entropy-based method can be extended to incorporate higher-order moments and more lag correlations, if needed. For example, in order to preserve kurtosis in the simulation, two Lagrange multipliers associated with the fourth noncentral moments of variables X and Y would be added in equation (3). Then, streamflow would be generated based on the corresponding conditional distribution as illustrated in section 2.5. Although the preservation of the kurtosis may not be essential and the sample instability problems with the estimation of higher moments may exist [*Fiering*, 1967], one simulation of this extension demonstrated the performance of the proposed method. Comparison of simulated kurtosis between the proposed method and the extended method, as shown in Figures 10(a) and 10(b), showed that the kurtosis was preserved better when the fourth moment was also incorporated as a constraint.

4.5. Drought, Surplus, and Storage Statistics

[50] Box plots of the drought, surplus and storage statistics (ratio of generated over historical) were constructed for simulation S_1 and S_2 and only the result for simulation S_2 is presented as shown in Figure 11. The water demand level was selected as a fraction of the historical mean and in this study it was selected as 0.7, 0.8, 0.9, and 1.0. For simulation S_2, as shown in Figure 11, the maximum deficit length and amount for the water demand level 1.0 were overestimated somewhat, but in general these statistics were preserved well. However, for simulation S_1 these statistics were not preserved as well.

5. Concluding Remarks

[51] A new model, based on entropy theory, for single-site monthly streamflow simulation is developed. Streamflow data is generated by sampling from the conditional distribution derived from the joint probability density function of streamflow of two adjacent months. The entropy-based model is applied to 10 sites in the Colorado River basin, and results indicate that it satisfactorily preserves mean, standard deviation, skewness, and lag-one correlation. Streamflow outside the observed streamflow range can be generated, though overestimation of the maximum values and underestimation of minimum values can occur when a relatively long record of annual streamflows is generated. Generally, drought, surplus, and storage statistics can be preserved well with the generation of a relatively long record.

[52] The advantage of the proposed method is that no assumption is made about the marginal distribution of the historical data. Therefore, the method can be applied to non-normal streamflow, and the transformation of streamflow to be normal is not needed. In addition, it can preserve the cross-correlation between streamflow in December (or last season) of the previous year and that in January (or first season) of the current year and avoid negative values in the generation. Further, if more statistical characteristics

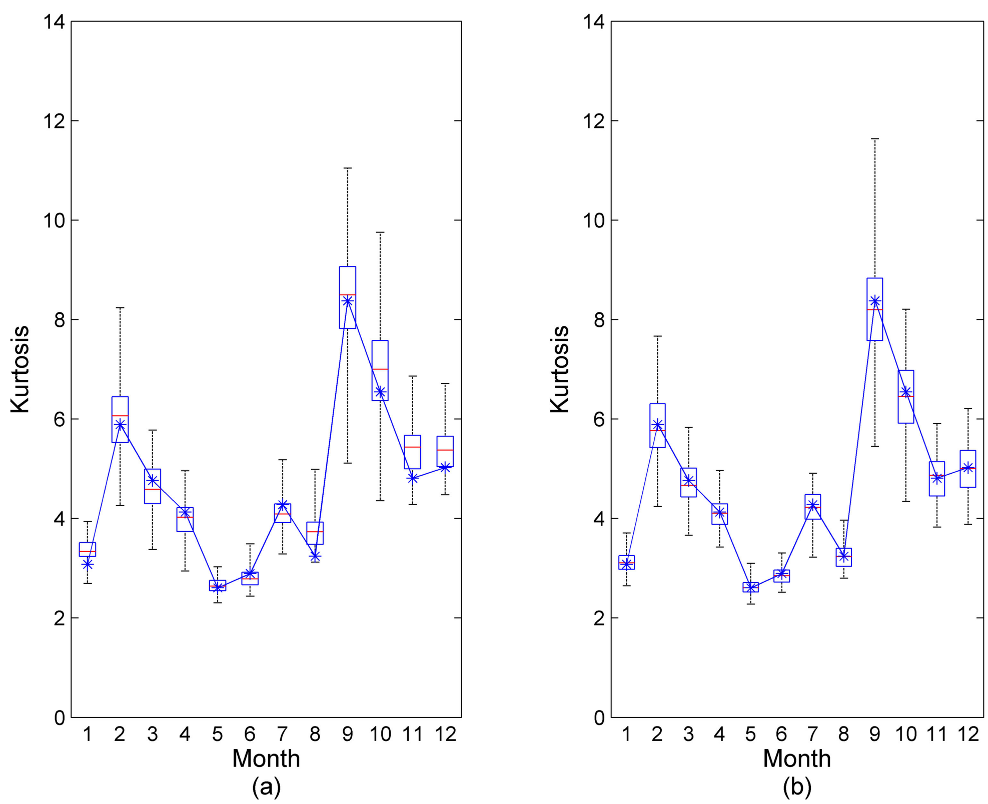

Figure 10. Box plots of kurtosis of historical and generated data for 100 sequences with 400 years of annual streamflow generated in each sequence. Continuous lines with star marks for each month represent historical data. (a) Proposed model; (b) extended model.

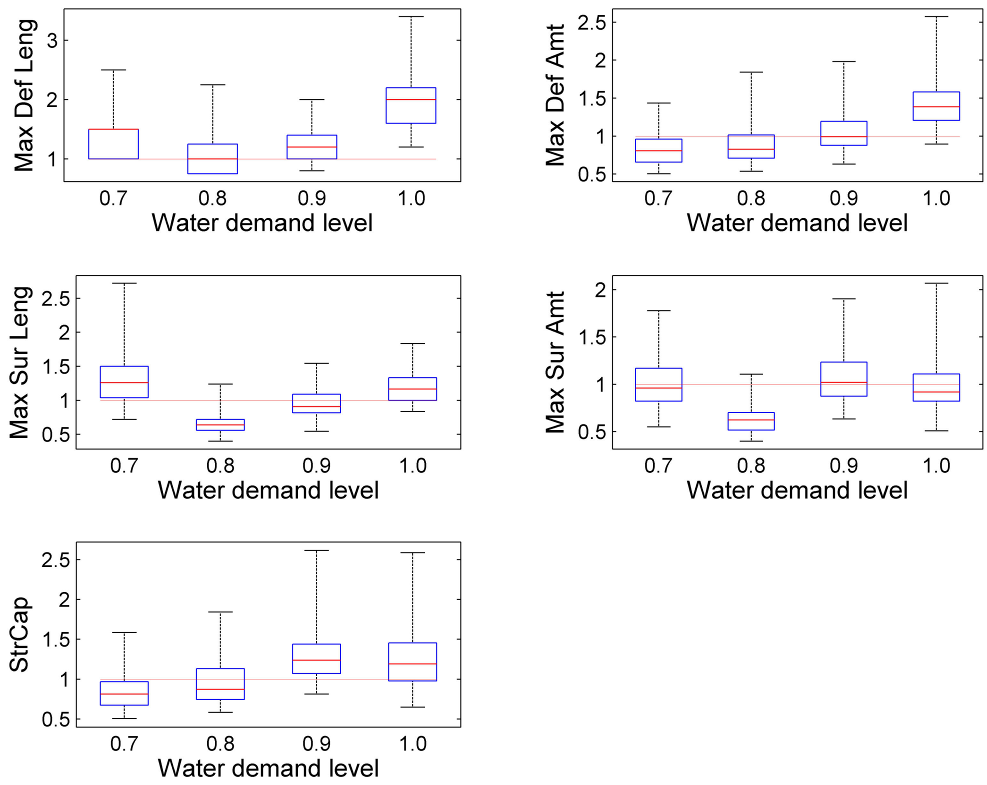

Figure 11. Box plots of ratio of drought, surplus and storage capacity statistics (for water demand level 0.7–1.0). Max Def Leng, maximum deficit length; Max Def Amt, maximum deficit amount; Max Sur Leng, maximum surplus length; Max Sur Amt, maximum surplus amount; StrCap, storage capacity.

(e.g., kurtosis and more lag correlations) are needed to be preserved, the entropy-based method can also be applied by incorporating these statistics as constraints. The disadvantage of the method is that it will be computationally cumbersome when more statistics are to be preserved and determination of more Lagrange multipliers is involved. This would be the case if the method were applied to multisite streamflow simulation, since statistics of streamflow at different stations would be used as constraints and integration in higher dimension will be involved in the determination of more Lagrange multipliers and streamflow simulation. However, this should not be an insurmountable difficultly, given the available numerical tools and computer progress. In addition, the bimodality that may exist in the empirical probability density function cannot yet be resolved with the proposed model.

[53] **Acknowledgments.** This work was financially supported in part by the United States Geological Survey (USGS, Project ID: 2009TX334G) and TWRI through the project "Hydrological Drought Characterization for TX under Climate Change, with Implications for Water Resources Planning and Management."

References

Balakrishnan, N., and C. Lai (2009), *Continuous Bivariate Distributions*, Springer, New York.

Fernandez, B., and J. D. Salas (1990), Gamma autoregressive models for stream flow simulation, *J. Hydraul. Eng.*, *116*(11), 1403–1414.

Fiering, M. (1967), *Streamflow Synthesis*, Harvard University Press, Cambridge, Mass.

Grygier, J., and J. Stedinger (1988), Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, *24*(10), 1574–1584, doi:10.1029/WR024i010p01574.

Hipel, K., and A. McLeod (1978), Preservation of the rescaled adjusted range, 2: Simulation studies using Box-Jenkins models, *Water Resour. Res.*, *14*(3), 509–516, doi:10.1029/WR014i003p00509.

Hipel, K., A. McLeod, and E. McBean (1979), Hydrologic generating model selection, *J. Water Resour. Plann. Manage. Div.*, *105*(2), 223–242.

Jaynes, E. (1957), Information Theory and Statistical Mechanics, *Phys. Rev.*, *106*(4), 620–630.

Johnson, M. (1987), *Multivariate Statistical Simulation*, John Wiley and Sons, New York.

Kapur, J. (1989), *Maximum-Entropy Models in Science and Engineering*, John Wiley, New York.

Kesavan, H., and J. Kapur (1992), *Entropy Optimization Principles with Applications*, Academic Press, New York.

Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*(7), 2105–2117, doi:10.1029/96WR00488.

Lall, U. (1995), Recent advance in nonparametric function estimation: Hydrologic application, *Rev. Geophys.*, *33*, 1093–1102, doi:10.1029/95RG00343.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*(3), 679–693, doi:10.1029/95WR02966.

Lee, T., and J. D. Salas (2006), Record extension of monthly flows for the Colorado River system, Bureau of Reclamation, U.S. Dept. of the Interior, Denver.

Lee, T., J. D. Salas, and J. Prairie (2010), An enhanced nonparametric streamflow disaggregation model with genetic algorithm, *Water Resour. Res. 46*, W08545, doi:10.1029/2009WR007761.

Lettenmaier, D., and S. Burges (1977), An operational approach to preserving skew in hydrologic models of long-term persistence, *Water Resour. Res. 13*(2), 281–290, doi:10.1029/WR013i002p00281.

Loucks, D., J. Stedinger, and D. Haith (1981), *Water Resource Systems Planning and Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

Mead, L., and N. Papanicolaou (1984), Maximum entropy in the problem of moments, *J. Math. Phys.*, *25*(8), 2404–2417.

Mejia, J., and J. Rousselle (1976), Disaggregation models in hydrology revisited, *Water Resour. Res.*, *12*(2), 185–186, doi:10.1029/WR012i002p00185.

Nowak, K., J. Prairie, B. Rajagopalan, and U. Lall (2010), A non-parametric stochastic approach for multisite disaggregation of annual to daily streamflow, *Water Resour. Res.*, *46*, W08529, doi:10.1029/2009WR008530.

Prairie, J., B. Rajagopalan, T. Fulp, and E. Zagona (2006), Modified K-NN model for stochastic streamflow simulation, *J. Hydrol. Eng.*, *11*(4), 371–378.

Prairie, J., B. Rajagopalan, U. Lall, and T. Fulp (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res.*, *43*, W03432, doi:10.1029/2005WR004721.

Salas, J., and J. Delleur (1980), *Applied Modeling of Hydrologic Time Series*, Water Resour. Publ., Littleton, Colo.

Salas, J., and T. Lee (2010), Nonparametric simulation of single site seasonal streamflows, *J. Hydrol. Eng.*, *15*(4), 284–296.

Santos, E., and J. Salas (1992), Stepwise disaggregation scheme for synthetic hydrology, *J. Hydrol. Eng.*, *118*(5), 765–784.

Savic, D., D. Burn, and Z. Zrinji (1989), Comparison of streamflow generation models for reservoir capacity-yield analysis, *Water Resour. Bull.*, *25*(5), 977–983.

Shannon, C. E. (1948), A mathematical theory of communications, *Bell Syst. Tech. J.*, *27*(7), 379–423.

Shannon, C., and W. Weaver (1949), *The Mathematical Theory of Communication*, Univ. Ill. Press, Urbana, Ill.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 1100, doi:10.1029/2001WR000953.

Sharma, A., D. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res. 33*(2), 291–308, doi:10.1029/96WR02839.

Smith, O., S. Adelfang, and J. Tubbs (1982), A bivariate gamma probability distribution with application to gust modeling, *NASA Tech. Mem. 82483*, Marshall Space Flight Center, Huntsville, Alabama.

Stedinger, J., D. Pei, and T. Cohn (1985), A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res.*, *21*(5), 665–675, doi:10.1029/WR021i1005p00665.

Tarboton, D., A. Sharma, and U. Lall (1998), Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res.*, *34*(1), 107–119, doi:10.1029/97WR02429.

Valencia, R. D., and J. J. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, *9*(3), 580–585, doi:10.1029/WR1009i1003p00580.

Vogel, R., and A. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resour. Res.*, *32*(6), 1875–1882, doi:10.1029/96WR00928.

Vogel, R., and J. Stedinger (1988), The value of stochastic streamflow models in overyear reservoir design applications, *Water Resour. Res.*, *24*(9), 1483–1490, doi:10.1029/WR1024i1009p01483.

Z. Hao and V. P. Singh, Department of Biological and Agricultural Engineering, Texas A&M University, 321 Scoates Hall, 2117 TAMU, College Station, TX 77843-2117, USA. (hzc07@tamu.edu; vsing@tamu.edu)